Evaluation of Spectral Tilt Measures for Sentence Prominence Under Different Noise Conditions

Sofoklis Kakouros[1], *Okko Räsänen*[1], *Paavo Alku*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland

sofoklis.kakouros@aalto.fi, okko.rasanen@aalto.fi, paavo.alku@aalto.fi

Abstract

Spectral tilt has been suggested to be a correlate of prominence in speech, although several studies have not replicated this empirically. This may be partially due to the lack of a standard method for tilt estimation from speech, rendering interpretations and comparisons between studies difficult. In addition, little is known about the performance of tilt estimators for prominence detection in the presence of noise. In this work, we investigate and compare several standard tilt measures on quantifying prominence in spoken Dutch and under different levels of additive noise. We also compare these measures with other acoustic correlates of prominence, namely, energy, F0, and duration. Our results provide further empirical support for the finding that tilt is a systematic correlate of prominence, at least in Dutch, even though energy, F0, and duration appear still to be more robust features for the task. In addition, our results show that there are notable differences between different tilt estimators in their ability to discriminate prominent words from non-prominent ones in different levels of noise.

Index Terms: prosody, sentence prominence, spectral tilt, noise, dnn

1. Introduction

Prosody and prosodic phenomena are critically important components of the spoken form of communication. In the same manner that paradigmatic contrasts at the segmental level allow the formation of linguistic units such as syllables or words, prosodic changes take place at slower rates and extend across individual segments conveying information about how something is spoken. Prominence is a prosodic phenomenon that can be generally defined as the property by which a linguistic unit is perceived to be standing out from its environment (see, e.g., [1,2,3], for related definitions). As the definition for prominence is domain specific, sentence prominence defines the degree of perceived emphasis for one or more words during a sentence (see, e.g., [2]). Prominence serves several functions in discourse, making it a particularly important component for natural language applications (see, e.g., [4,5]). For instance, prominence can be indicative of the lexical class or information structure in a sentence [3]. In this regard, prominence may cue the most important word or words in a sentence and therefore reflect the communicative intentions of the speaker through emphasis.

The study of prominence has indicated a number of factors that seem to hold a role in the production and perception of prominent units in speech. Purely from the physical signal perspective, four acoustic correlates for prominence have been identified across a number of studies. Specifically, energy [6,7,8], fundamental frequency (F0) [9], duration [7,8], and spectral tilt [10,11,12] have been all observed to correlate with the incidence of prominent units in speech. However, the independent contribution of energy, F0, and duration in prominence seems to be more established across studies than that of spectral tilt. This might be explained by the fact that there are more established and reliable measures for energy, F0, and duration than for spectral tilt. For spectral tilt, a variety of measures have been proposed in the literature, often encountered under different terms.

The diversity of the measures quantifying spectral tilt may also pose challenges in the interpretation of results across different studies. For instance, Sluijter and van Heuven measured spectral tilt as the band-limited intensity across four continuous spectral bands (0–0.5, 0.5–1, 1–2, and 2–4 kHz) [10]. In another study, Campbell and Beckman used the harmonic ratio (difference in dB between the first and second harmonic of F0, H1-H2) in order to quantify a measure for spectral tilt [11]. Other studies use an array of different methods, including calculation of the difference in dB between the overall intensity and the intensity of the fundamental frequency (or in a frequency band centered at the fundamental) [12,13,14], taking the first cepstral coefficient [15], taking the difference in dB between a signal with high-frequency pre-emphasis and flat frequency weighting (SPHL-SPL) [16], taking the difference in dB between the first harmonic and third formant (H1-F3) [17], fitting a regression line in the magnitude spectrum [18,19], taking the band-limited spectral energy ratios [20,21], using the long-term average spectrum (LTAS) to obtain band-limited energy ratios [22], and using all-pole modeling techniques [23]. In addition, some studies utilize measures such as regression line fitting and harmonic ratio, but, instead of applying the measures directly on the short-term spectrum of speech, they utilize the spectrum of the glottal source waveform that is obtained through glottal inverse filtering (GIF) [24,25]. Recently, a DNN-based system for robust spectral tilt estimation was described in [26], enabling the estimation of the glottal source tilt in non-ideal signal conditions without explicitly using GIF in the estimation phase.

As the methods for the estimation of spectral tilt are highly variable, the goal of the present work is to compare a range of the most well-known measures for spectral tilt together with the newly-proposed deep neural net (DNN) –based technique [26]. The capability of the spectral tilt measures to describe prosodic prominence is then evaluated using clean and noise-corrupted Dutch speech. The results indicate that tilt is a consistent correlate of prominence, but also that there are important differences in the discriminative capabilities of the different tilt estimators.

2. Methods

Central part of the investigation in this work is the evaluation of the effect of noise across different features for prominence. For this purpose, speech signals were initially degraded with additive babble noise of variable signal-to-noise ratio (SNR). For each signal, the standard acoustic correlates for prominence, namely, energy, F0, and duration were computed (section 2.1) together with several tilt measures that have been commonly used in the literature (section 2.2). In addition, a DNN-based spectral tilt estimation was also evaluated (similar to [26]) in order to investigate (i) the efficiency of DNN-based source tilt estimation for prominence, and (ii) the potential for the DNN to add robustness on tilt estimation for noisy signals (section 2.3). For all features, a number of aggregate statistical measures over words were then computed, and their capability to discriminate prominent from non-prominent words was measured in terms of the separability of the feature distributions. All measures were computed only for the voiced frames during the words, as detected by the F0 estimator (see below) at each given noise level.

2.1. Energy, F0, and duration

Energy, F0, and duration were used as the reference features in this work because many previous studies have shown that they correlate well with the manifestation of prominence (see, e.g., [6,7,8,9,10]). In order to compute them, speech data were initially downsampled to 16 kHz. F0 estimation was carried out using a noise robust pitch tracker [27] with a 100-ms window and 10-ms hop size. The pitch tracker provided pitch estimates as well as a voicing decision for each frame of the analysis. Finally, energy was computed using a 20-ms window and 10-ms hop size, while duration was extracted for each word from the corpus annotations.

2.2. Spectral tilt measures

For the comparative analysis of the spectral tilt measures, several tilt estimation techniques that are commonly used in the literature were utilized. In this work we are limiting the analysis to include only scalar one-parameter models. All tilt measures below were computed over a 20-ms window and using a 10-ms hop size. The following measures were included in the comparison:

- The difference in dB between the first and second harmonic (H1-H2) (see, e.g., [11]).
- The difference in dB between the first harmonic and third formant (H1-F3) (see, e.g., [17]).
- The first cepstral coefficient (see, e.g., [15]).
- The spectral energy ratio in the frequency bands between 0–1 kHz and 1–5 kHz (see, e.g., [20]).
- The slope of the line obtained by fitting a first order polynomial in the short-term magnitude spectrum (spectral regression – see, e.g., [18]).
- The first order linear prediction coefficient (LPC) – 1LP.

2.3. DNN-based spectral tilt estimation

A new method was proposed recently in [26] to estimate and parameterize the glottal source spectrum in noisy, non-ideal conditions where conventional GIF analysis cannot be used due to its known sensitivity to noise [28]. The method proposed in [26] uses a deep neural network (DNN) to map an input feature vector (the logarithmic speech power spectrum) into an output vector (all-pole model of the glottal source spectrum parameterized using line spectrum frequencies (LSFs)). In this work, two separate standard feed-forward DNNs were trained for the prediction of the LSFs describing the glottal source spectrum directly from the logarithmic FFT magnitude spectrum of the speech input (20-ms window, 10-ms hop size; see also Fig. 1). The first DNN was trained on clean speech only. In the second DNN the clean training data were augmented with noise-corrupted versions of the same data with 15 dB SNR additive babble noise but using the clean speech LSFs. It is worth emphasizing that the DNN-based spectral tilt estimation method does not need GIF in the estimation phase. GIF is used only in the training phase to compute the output LSF feature vectors from studio-quality speech. In the current study, the quasi closed phase (QCP) method [29] was used in the computation of the output LSF feature vectors.

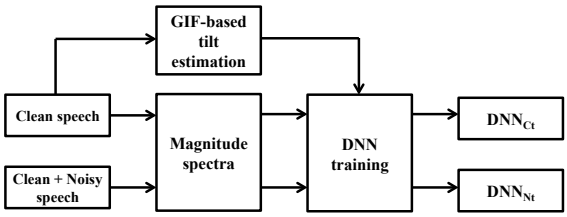

Figure 1: *Schematic diagram of the training of the DNN-based tilt estimator.*

The 255-dimensional spectral frame inputs and LSF outputs of the DNNs were z-score normalized across all training data to ensure proper scaling. Both networks use sigmoid activation functions for hidden layers, a linear output layer, a learning rate of $\eta = 0.1$, 100 epochs, minibatch size of 1000, mean squared error (MSE) as the cost function, and a configuration layout d = [64 32 16] for the hidden units per layer. This results in two DNNs for tilt prediction: one based on clean speech (DNN_{Ct}) and a second based on a clean together with noise-corrupted speech (DNN_{Nt}). The final tilt estimates are then obtained by fitting a first order polynomial in the spectrum of the glottal waveform as parametrized by the predicted LSFs. We also compare the resulting tilt estimates to those computed directly from speech using normal QCP.

3. Experiments

3.1. Material

Two speech corpora were used in our study. The Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) was used as the basis for the evaluations of the different acoustic measures. CGN is a database of contemporary standard Dutch as spoken by adults in the Netherlands and Flanders (see [30] for more details). For the present evaluations, the prosodically annotated subset of the Dutch news broadcast (*component-k*) part of the corpus containing prominence annotations was utilized, consisting of 134 news broadcasts spoken by 10 different speakers (9 male and 1 female) (≈44.3 minutes of speech data). The annotations were hand-labeled using binary (prominent/non-prominent) markings by two trained annotators, containing a total of 7438 word tokens.

A second corpus was used to provide high quality clean speech signals for the DNN training. For this purpose, the Phonetic Corpus of Estonian Spontaneous Speech of the University of Tartu was utilized (“EstPhon”) [31]. The corpus contains a total of 60 hours of conversational recordings by speakers from different age groups, dialectological, and social backgrounds. In this work, we used 1165 randomly chosen

utterances from the high-quality studio section of the corpus to train the DNNs for QCP tilt estimation.

3.2. Evaluation

All evaluations in this study were carried out at the word level on the CGN corpus. Specifically, the manually labeled prominence markings were used to divide the data into two categories: prominent and non-prominent words. As the data have been labeled by two annotators, all words with at least one prominence marking were considered as prominent (see [32] for a similar approach). For the evaluation, five word-level statistical descriptors were computed for all measured features: (i) mean, (ii) max, (iii) min, (iv) standard deviation (SD), and (v) the feature range during the word.

In order to compare the differences in feature distributions for the prominent and non-prominent classes, the estimated Z-score based effect-size r from Wilcoxon rank sum test (Eq. (1)) was utilized together with the symmetric Kullback–Leibler (KL) divergence (Eq. (2)). KL-divergence was computed using $Q = 25$ discrete bins with all having a uniform number of samples across the entire data set. Both measures quantify the degree of separability of the prominent and non-prominent classes with zero corresponding to no difference. In the equations, P_{pr} and P_{npr} denote the probability density of the matching bins and N_{pr} and N_{npr} denote the number of samples for the two classes, respectively.

$$r = \frac{Z}{\sqrt{N_{pr} + N_{npr}}} \tag{1}$$

$$D_{KL} = \sum P_{pr} \log\left(\frac{P_{pr}}{P_{npr}}\right) + \sum P_{npr} \log\left(\frac{P_{npr}}{P_{pr}}\right) \tag{2}$$

4. Results

After training the DNNs (DNN_{Ct} and DNN_{Nt}) using the Estonian corpus, the annotated part of the CGN *component-k* was used to compute all features from both clean and noisy versions of the 134 speech signals. The noisy versions of the CGN signals were generated by corrupting them with additive babble noise (different from the one used for corrupting the DNN training data) with SNRs of -10, -5, 0, 5, 10, 15, 20, 40, and 60 dB in addition to using clean speech. It is important to note here that the CGN is inherently noisier than the EstPhon data, and therefore "clean" refers to the potentially non-ideal signal quality in the broadcast recordings. Since the overall behavior of D_{KL} was found to be nearly identical to the effect size r across all conditions, we only report the latter in the following sections.

4.1. Energy, F0, and duration

Fig. 2 shows the results for energy and F0. As expected, the class separation between the two prominence classes is robust. Substantial variation between the different statistical measures can be also observed, and the measure for min is not separately shown for any of the measures as it performed generally poorly. The highest class separation is attained for the SD and range whereas the lowest for max and mean. In addition, range, SD, and max seem to maintain class separability across the entire tested SNR range while mean drops to near zero for low SNRs. The non-zero separability for low SNRs is likely due to the confounding effect of duration where longer words are intrinsically more likely to exhibit a wider range of values. The durational measure is not plotted in Fig. 2 as it is independent of SNR, achieving a class separation

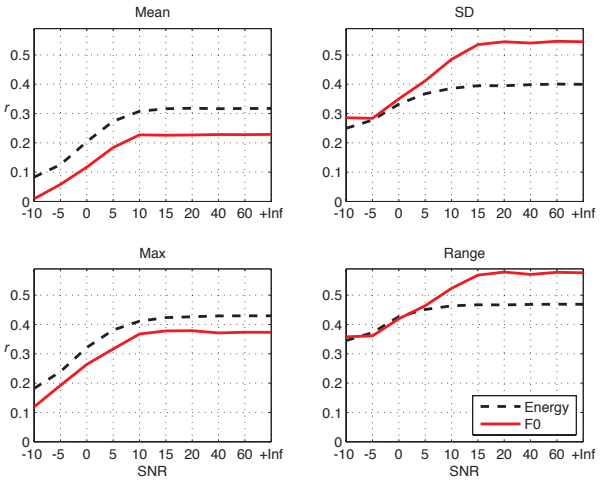

Figure 2: *Prominent and non-prominent class separations (r) for energy and F0 plotted for mean, SD, max, and range and from -10 to 60 dB SNR and clean signals.*

of $r = 0.72$. In all, all three features have strong descriptive capabilities in the task of discriminating between the prominent and non-prominent classes.

4.2. Spectral tilt measures

A total of nine tilt measures are evaluated in this work. The results for the spectral tilt measures are divided into three subsections for clarity of presentation.

4.2.1. H1-H2, H1-F3, and 1LP

The first set of tilt measures can be seen in Fig. 3 (H1-H2, H1-F3, 1LP). Similar to the case for energy and F0, the descriptors for range and SD provide the best overall separability across all measures. However, contrary to energy and F0, the estimates converge to zero for low SNRs (< 5 dB) for all descriptors. It can be also noticed from Fig. 3 that the features behave slightly differently across the different statistical descriptors. Specifically, whereas 1LP appears to provide the best class separation for mean, max, and SD, this does not hold for the range where H1-F3 and H1-H2 are better. Overall, the best performing feature in this group is H1-F3 using range as the descriptor ($r = 0.43$ for clean speech).

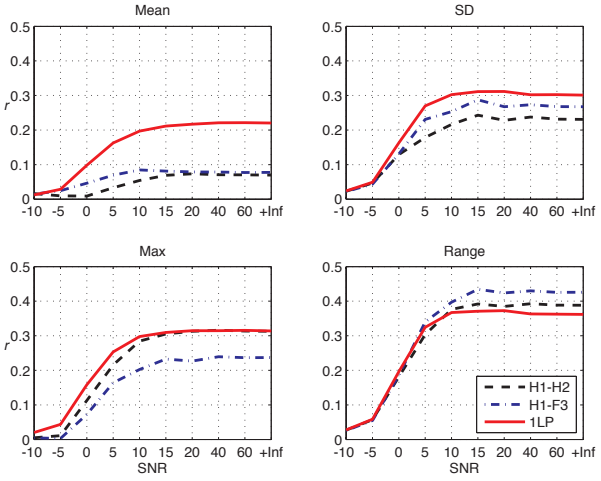

Figure 3: *Prominent and non-prominent class separation (r) for H1-H2, H1-F3, and 1LP plotted for mean, SD, max, and range and from -10 to 60 dB SNR and clean signals.*

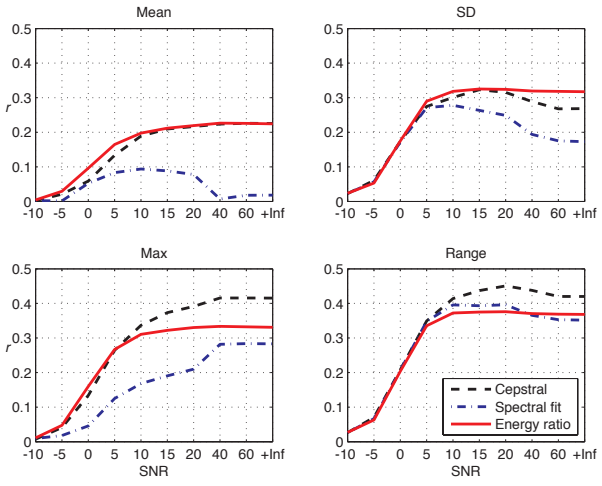

Figure 4: *Prominent and non-prominent class separation (r) for first cepstral coefficient, regression line fit, and energy ratio plotted for mean, SD, max, and range and from -10 to 60 dB SNR and clean signals.*

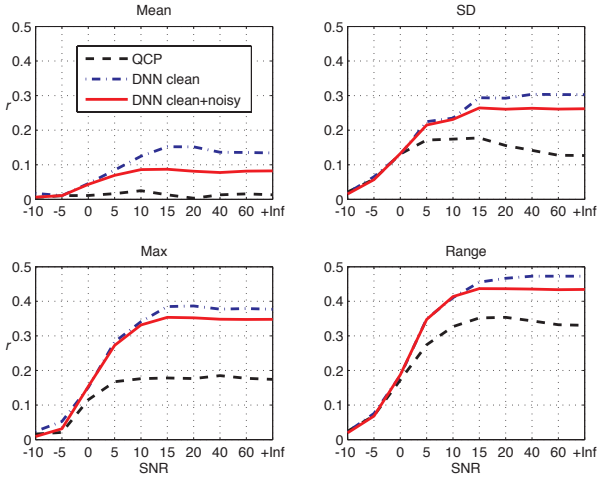

Figure 5: *Prominent and non-prominent class separation (r) for QCP, DNN_{Ct} (clean), and DNN_{Nt} (clean+noisy) plotted for mean, SD, max, and range and from -10 to 60 dB SNR and clean signals.*

4.2.2. Cepstral, regression line slope, and energy ratio

The performance of the second analysis group can be seen in Fig. 4. As with the previous tilt measures (section 4.2.1), in this case too, there is substantial variation in the performance across different statistical descriptors. Between the three measures, the spectral fit slope appears to be the most volatile and also the worst performing feature for mean, SD, and max. Here too, the best performance is attained for range and for the cepstral tilt estimate ($r = 0.42$ for clean speech).

4.2.3. DNN-based source tilts and direct QCP-based tilt

The performance of the DNN-based tilt estimates together with the direct QCP tilt estimation can be seen in Fig. 5. For SD, max, and range, the performance differences between the methods can be observed where the DNN-based tilt estimates outperform the direct QCP-based estimation. Overall, the DNN-based approach provides more robust estimates for tilt with increasing noise level than the direct QCP on the same data. Here too, the best performance was obtained for feature range for both DNNs (DNN_{Ct} and DNN_{Nt}).

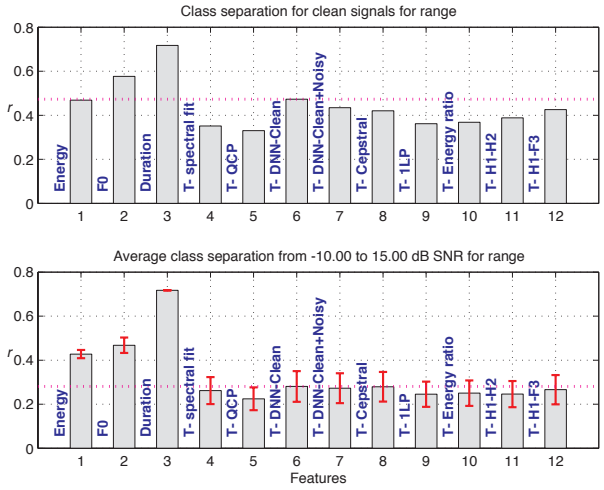

Figure 6: *Prominent and non-prominent class separation (r) for all evaluated features for the range descriptor. Top: clean signals. Bottom: averaged r for SNRs between -10 to 15 dB where vertical red bars denote the standard error.*

5. Discussion and Conclusions

The analysis of the tilt estimates provided insights into the behaviour of the measures for prominence. Specifically, it was observed (i) that the tilt measures behave differently for the distinct statistical descriptors, (ii) there are differences in the performance (class separability) for the distinct tilt measures in clean speech, and (iii) noise degradation greatly impacts prominence class separation from around 10 dB SNR, largely diminishing the differences between the estimators. The results also support the finding that tilt is an important correlate for prominence, at least for Dutch (see, e.g., [10,33], see also [34]).

Fig. 6 presents a comparative overview of all the features evaluated in this study. It can be seen that energy, F0, and duration are the most robust across all features. Across the tilt measures and for the clean speech signals, both DNN-based tilt estimators seem to perform best followed by H1-F3 and the first cepstral coefficient. With additive noise, the best performing tilt estimates remain the DNNs and cepstral tilt. In addition, another finding from the current evaluation is that, at least for tilt, the statistical descriptor that offers the most discriminative power in the case of prominence is the range of the feature (max-min) during each word. With respect to the DNN-based tilt estimation, it was shown that the DNN can provide better tilt estimates for the task improving over a direct GIF estimation of the tilt from the speech signal. One limitation of the present evaluation is that all measures were examined over voiced frames (over words), thereby intrinsically binding them to the F0 estimation procedure. This means that for decreasing SNR, the potential number of voiced frames will reduce, inevitably also reducing the number of samples for the measures. In future investigations, the measures should be also computed for each frame over the entire word or using the clean speech voicing estimates for the noisy versions of the signals.

6. Acknowledgements

This study was funded by the Academy of Finland (project no. 274479 and 284671).

7. References

[1] J. Terken and D. Hermes, "The perception of prosodic prominence," in *Prosody: Theory and experiment. Studies presented to Gösta Bruce,* M. Horne, Ed. Dordrecht, The Netherlands: Kluwer, pp. 89–127, 2000.

[2] A. Cutler, "Lexical Stress," in *The handbook of speech perception*, D. B. Pisoni and R. E. Remez, Eds. Blackwell publishing, pp. 264–289, 2005.

[3] P. Wagner et al., "Different parts of the same elephant: a roadmap to disentangle and connect different perspectives on prosodic prominence," in *Proceedings of ICPhS*, 2015.

[4] M. Mehrabani, T. Mishra, and A. Conkie, "Unsupervised prominence prediction for speech synthesis," in *Proceedings of INTERSPEECH,* pp. 1559–1563, 2013.

[5] D. N. Racca and G. J. Jones, "Incorporating Prosodic Prominence Evidence into Term Weights for Spoken Content Retrieval," in *Proceedings of INTERSPEECH,* pp. 1378–1382, 2015.

[6] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1038–1054, 2005.

[7] D. B. Fry, "Duration and intensity as physical correlates of linguistic stress," *The Journal of the Acoustical Society of America*, 27(4), pp. 765–768, 1955.

[8] P. Lieberman, "Some acoustic correlates of word stress in American English," *Journal of the Acoustical Society of America*, vol. 32, no. 4, pp. 451–454, 1960.

[9] J. Terken, "Fundamental frequency and perceived prominence of accented syllables," *Journal of the Acoustical Society of America*, vol. 89, no. 4, pp. 1768–1776, 1991.

[10] A. M. C. Sluijter and V. J. van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2471–2485, 1996.

[11] N. Campbell and M. E. Beckman, "Stress, prominence, and spectral tilt," In Botinis, A., Kouroupetroglou, G., and Carayiannis, G. (Eds.), Intonation: Theory, Models, and Applications (Proceedings of an ESCA Workshop), pp. 67–70, 1997.

[12] M. Heldner, "Spectral emphasis as an additional source of information in accent detection," in *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, 2001.

[13] P. A. Barbosa, A., Eriksson, and J. Akesson, "On the Robustness of some Acoustic Parameters for Signaling Word Stress across Styles in Brazilian Portuguese," in *Proceedings of Interspeech*, pp. 282–286, 2013.

[14] A. Eriksson, G. C. Thunberg, and H. Traunmüller, "Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing," in *Proceedings of EuroSpeech*, pp. 399–402, 2001.

[15] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Spectral moment features augmented by low order cepstral coefficients for robust ASR," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 551–554, 2010.

[16] G. Fant, A. Kruckenberg, J. Liljecrants, and S. Hertegård, "Acoustic-phonetic studies of prominence in Swedish," TMH-QPSR, vol. 2–3, pp. 1–51, 2000.

[17] A. O. Okobi, "Acoustic correlates of word stress in American English," Dissertation, Cornell University, 2006.

[18] Aronov, G. and Schweitzer, A., "Acoustic correlates of word stress in German spontaneous speech," in *Proceedings of Tagung Phonetik und Phonologie im Deutschsprachigen Raum*, 2016.

[19] Y. Lu and M. Cooke. "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. *51*(12), pp. 1253–1262, 2009.

[20] P. J. Murphy, K. G. McGuigan, M.Walsh, and M. Colreavy, "Investigation of a glottal related harmonics-to-noise ratio and spectral tilt as indicators of glottal noise in synthesized and human voice signals," *The Journal of the Acoustical Society of America*, vol. *123*(3), 1642–1652, 2008.

[21] P. Prieto and M. Ortega-Llebaria, "Stress and Accent in Catalan and Spanish: Patterns of duration, vowel quality, overall intensity, and spectral balance," in *Proceedings of Speech Prosody*, pp. 337–340, 2006.

[22] J. Sundberg and M. Nordenberg, "Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech," *The Journal of the Acoustical Society of America*, vol. *120*(1), pp. 453–457, 2006.

[23] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. *51*(5), pp. 401–411, 2009.

[24] M. Jackson, P. Ladefoged, M. Huffman, and N. Antoñanzas-Barroso, "Measures of spectral tilt," *The Journal of the Acoustical Society of America*, vol. *77*(S1), p. S86, 1985.

[25] J. Kreiman, B. R. Gerratt, and N. Antonanzas-Barroso, N., "Measures of the glottal source spectrum," *Journal of speech, language, and hearing research*, vol. *50*(3), pp. 595–610, 2007.

[26] E. Jokinen and P. Alku, "Estimating the spectral tilt of the glottal source from telephone speech using a deep neural network," *The Journal of the Acoustical Society of America,* accepted for publication, 2017.

[27] T. Drugman and A. Alwan, "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics," in *Proceedings of Interspeech,* pp. 1973–1976, 2011.

[28] P. Alku, "Glottal inverse filtering analysis of human voice production – A review of estimation and parameterization methods of the glottal excitation and their applications," Sadhana – Academy Proceedings in Engineering Sciences, vol. 36, part 5, pp. 623–650, 2011.

[29] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. *22*(3), pp. 596–607, 2014.

[30] J. Duchateau, T. Ceyssens, and H. Van Hamme, "Use and evaluation of prosodic annotations in Dutch," in *Proceedings of LREC,* pp. 1517–1520, 2004.

[31] Phonetic Corpus of Estonian Spontaneous Speech, Retrieved from http://www.keel.ut.ee/en/languages-resourceslanguages-resources/phonetic-corpus-estonian-spontaneous-speech, 2017.

[32] S. Kakouros and O. Räsänen, "3PRO–An unsupervised method for the automatic detection of sentence prominence in speech," *Speech Communication*, vol. *82*, pp. 67–84, 2016.

[33] S. Kakouros, J. Pelemans, L. Verwimp, P. Wambacq, and O. Räsänen, O., "Analyzing the Contribution of Top-down Lexical and Bottom-up Acoustic Cues in the Detection of Sentence Prominence," in Proceedings of *Interspeech*, pp. 1074–1078, 2016.

[34] B. M. Streefkerk, L. C. Pols, and L. ten Bosch, "Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANNs," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH-1999)*, pp. 551–554, 1999.